S.A. Shahab, Wasfi G. Al-Khatib, **Sabri A. Mahmoud**,
"Computer Aided Indexing of Historical Manuscripts."
The International Conference, Computer Graphics, Imaging and Visualization, 25-28 July 2006, Sydney, Australia.

Key words: Document image analysis, historical manuscripts, Arabic manuscripts, content-based retrieval, similarity measures.

Computer Aided Indexing of Historical Manuscripts

S.A. Shahab, Wasfi G. Al-Khatib, Sabri A. Mahmoud
Information and Computer Science Department
King Fahd University of Petroleum & Minerals
Dhahran 31261, Saudi Arabia
{sadnans,wasfi,smasaad}@ccse.kfupm.edu.sa

Abstract

Arabic manuscripts represent a rich source of knowledge that has been highly underutilized. Huge repositories of historical artifacts are yet to be typeset and published in book-form. Given vast content of these manuscripts, it is important to develop indexing systems that support content-based retrieval from historical manuscripts. In this paper, we propose a computer aided retrieval and indexing system for Arabic historical manuscripts. The proposed system extracts meaningful information (features) that is used in indexing. Some preprocessing steps are also implemented in order to enhance the quality of document images. More than one form of a similarity measure has been tested. The developed prototype system has shown encouraging results with respect to the word matching rates achieved.

Keywords— **Document image analysis, historical manuscripts, Arabic manuscripts, content-based retrieval, similarity measures.**

83%. Arabic is no exception. The Arabic writing is cursive. Since character shapes change pending their position in a word, word segmentation into characters is a major problem.

In our work, we shall address the issue of content based retrieval of historical Arabic manuscripts. This paper is organized as follows: Section 2 presents other related work to our work. Then, we describe our proposed framework of the system in Section 3. Section 4 presents the experimental results, followed by the conclusion.

2 Related Work

Different techniques for various aspects of document image analysis and retrieval have been developed. For example, digitization of documents that have not been generated from a computer, such as those of manuals and documents that have been typeset, has been mainly based on applying information retrieval techniques to text that has been recognized through OCR methods [3, 4]. Work in this area focuses on handling the erroneous character recogni-